

Palick, L. K
Pt. 2
C-1

Meteorologic Impacts of Forest Management Activities
on Wildfire Potential

Part 2 - Procedures for Extension of Meteorological
Observations to Remote Sites

LIBRARY

JUN 1 1978

ROCKY MOUNTAIN STATION

FINAL REPORT

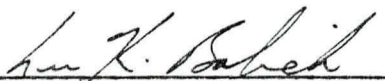
Cooperative Research Agreement
CSU Project No. 31-1470-2184
RMF&R Experiment Station No. 16-639-CA

Meteorologic Impacts of Forest Management Activities
on Wildfire Potential

Part 2 - Procedures for Extension of Meteorological
Observations to Remote Sites

(Part 1 - Technical Report is Under Separate Cover)

Prepared by


Lee K. Balick

Jack S. Barrows

LIBRARY

JUN 1 1978

ROCKY MOUNTAIN STATION

ABSTRACT

A procedure to develop regression equations for meteorological observations between two locations is described. The procedure uses one year of observations at a remote site and a permanent or "base" weather station to derive a set of regression equations. Equations can then be applied to the long term records of the base station to estimate climate at the remote site. The procedure is designed to give improved site specific estimates of climatological variables that are inputs to computer models of wildfire hazard.

Discussion is presented on two levels. First the reasoning and philosophy of the steps is given. Then the application of the reasoning is applied to an example. A user must make some decisions in applying the technique to individual management problems. Recommendations based on regression concepts and experience gained in developing the regression procedure are presented.

Computational steps are designed for and discussed in terms of the Statistical Package for the Social Sciences. This is an easily learned software package designed for use by non-statisticians.

Meteorologic Impacts of Forest Management Activities on Wildfire Potential

Part 2 - Procedures for Extension of Meteorological Observations to Remote Sites

1. Introduction

This report presents a technique for making site specific estimates of climatic variables at locations where no long term records exists. One year of observations at the remote site are used with measurements of the same year at a nearby permanent weather station to develop a set of regression equations. The relationships between the two sites as contained in the regression equations are then applied to the historical records of the permanent station. An estimate of the climate at the remote station is then obtained. Technical details and testing of the procedure are discussed in Part 1 under separate cover. This part presents the step by step application of the technique.

To apply the technique a computer is required. Programming skills and statistical decision making are reduced as much as possible. To do this the Statistical Package for the Social Sciences (SPSS) is used. This software package is easy to learn and use and is available on almost all large computers. The documentation (Nie et. al. 1975) contains a thorough description of what the programs do and what the numbers mean. It is recommended that the reader have a copy of the SPSS manual at hand. The nomenclature used in this report is consistent with SPSS documentation. On many computers the SPSS version described by Nie et. al. (1975) has been updated. Updates include a full stepwise multivariate regression program that is not in the 1975 version.

(On Control Data Corporation computers the update is SPSS Version 6.5.

Users should obtain supplemental documentation of the particular version on the computer they will use. If the version available to the user is the 1975 version, then forward selection may be used as opposed to the stepwise technique used here.) Every application and every data set has its own unique features. It is not possible or even wise to put the regression procedure described here entirely within a "block box." An attempt has been made to simplify the procedure as much as possible and to minimize the decisions required.

Instrumentation at the remote site is expected to be limited to a hygromograph and a recording air movement meter or their equivalents. Therefore the variables that are to be estimated (dependent variables) are maximum and minimum temperature, maximum and minimum relative humidity and wind speed. Other variables that are inputs to fire hazard models are physically discrete (and therefore intractable; precipitation) or require an observer or expensive instrumentation. Early in the development of the regression procedure, it was found that relative humidity is a difficult variable to use. Its frequency distributions are assymetrical and bounded between 0 and 100. Also, relative humidity is a strong function of temperature. It is best to maintain independence of the variables and to use variables that have conservative properties. A number of unbounded and quasi-conservative substitutes for relative humidity are available: dew point, vapor pressure and specific

humidity are among them. Dew point was chosen for the work presented here because of its general familiarity and it was already established that it had a symmetrical distribution (Barrows and Balick, 1977). A synthetic afternoon dew point is created from maximum temperature and minimum relative humidity and a morning dew point is calculated from minimum temperature and maximum relative humidity. The relationship between temperature, relative humidity and dew point is given in Appendix A.

Discussion of the regression procedure is presented on two levels. Because the user must make decisions the reasoning behind each step is presented. Then the reasoning is applied to an example. The same example is developed throughout the text and is hypothetical (but based on work presented in Part 1). The example cannot cover every situation so additional alternatives, trouble spots and other details are discussed.

The regression procedure is divided into three phases. The first is a problem overview and set-up. This phase includes all steps from problem identification to creation of the data files. Next, the set of terms for possible inclusion in the regression equations is constructed. In the third step a regression equation for each dependent variable is developed. After development the equations can be applied to historical data.

2. Procedure Set-up

This initial phase involves all steps up to the point where calculations begin. It starts with the user becoming familiar with the climate of the area and ends with the preparation of a computer data file.

There are several reasons for the user becoming as familiar as possible with the climate of both the general area and the specific locality where the procedure is to be applied. First, several steps involve subjective

decisions which are best made with good information. Each area or each data set likely has its unique features to which this procedure must be adapted. Some of these can be anticipated if the user has knowledge of the area. Lastly, there is not quantitative check on the accuracy of the regression equations when they are applied to other years. It is important that they be found to give reasonable estimates. "Reasonableness" must be judged in terms of what is expected for the area. Objective criteria for guiding decisions are established but the final decision must be subjective. If initial results seem unsatisfactory, then changes or adaptations of the procedure should be guided by knowledge of the climate.

There are three decisions that must be made before observations begin. The first is choosing the location for the remote site weather station. Next, the beginning and ending dates for operation of the remote station are chosen. Finally, the fire weather station or base station is selected.

The remote weather station should be placed in an area that best represents the entire site to be studied. It should not be placed in local depressions or on knolls. If wind speeds are to be estimated it should not be where two drainages join. If the site is being considered for management treatments that affect stand density, it is very important that the wind speed meter be located in a clearing. Wind speed under a canopy is very low and this presents two problems. The first is that most anemometers are not accurate at low wind speeds. Low wind speeds also reduce the chance to obtain good statistical relationships. Under a canopy wind observations may be nearly constant, in a statistical sense, and there is no correlation between a constant and a variable. Not only do poor regression relationships result from this situation but it is impossible to estimate

the effects of changes of stand density on wind speed (see Part 1).

The remote station should be operated during the entire length of each fire season of the year. Since the beginning and end of a season change from year to year, it is best to plan to start observations a little early and end late and later eliminate data that do not represent the season. It is difficult to specify the minimum length of a season. It seems that 50 to 60 days is adequate for seasons with relatively stable weather but 70 to 80 days is needed when a larger variety or rapidly changing meteorological conditions exist. In later steps it may be necessary to redefine seasons. If so, some extra days data may provide valuable flexibility. Finally, the remote station need be operated only while the base station is in operation.

Conceptually, the base station can be selected after the season but it is advisable to choose it as early as possible. If selection is made before the season, then effort can be applied toward maximizing data quality. Arrangements can be made for extra training of observers, increased frequency of instrument calibration and assure that observations are made every day. Data quality, both at the remote and base stations, is very important. There are two main criteria for selecting the base station. It should have a relatively long record of complete observation. The regression equations will be applied to the period of record at that station. The second consideration is that the base station should have the highest expected correlation with the remote station of all possible choices. To the extent possible it should be located in the same drainage or local air circulation as the remote site. Distance is an important factor that should be minimized but not for all situations. A mountain barrier between two stations may result in a larger loss of correlation than several miles of distance. The choice should be

based on knowledge of the area and often consultation with local people is beneficial.

Once the observations are made the task is to put them on a computer file in an SPSS compatible format. The goal is to have all observations from both stations on one record for each day. (At some stage relative humidities are transformed to dew point values). There are many ways to do this and the choice is dependent on the facilities and skills available to the user. Once this is done the data are then screened for errors. If the errors are of the transcription or keypunch type, they can be corrected. If there are still apparently bad data points, then the source of error (if any) must be further investigated and decisions must be made for their disposition. When a clean data set is produced, then calculations can begin.

At this point we begin developing the example which is followed throughout the remainder of this report. The base station is the fire weather station at Flagstaff, Arizona and the remote station is in a clearing in ponderosa pine south of the base station. Wind speed observations are not available at the remote station so only daily maximum and minimum temperature and afternoon and morning dew points are to be estimated. (Symbolism and notation other than that of SPSS are given in Appendix A. The conversion between relative humidity and dew point is given in Appendix B).

We start the example where the records from Flagstaff have been merged with those of the remote station and dew points are computed. The SPSS variable list and input format cards are

```
VARIABLE LIST  YR,MO,DY,SW,TX,TN,ADP,MDP,PA,YTX,YTN,YADP,YMDP
INPUT FORMAT   FIXED (2X,3F2.0,F1.0,2F3.0,3F4.0,2F3.0,2F4.0)
```

Columns 1 and 2 contain a station identification which is not used. Only six variables from the base station are used; the others are judged to be inappropriate, unreliable or too intermittent after screening of the Flagstaff data. Missing values are set to -99. An eighty day season runs from April 20 through June 30. (In reality two seasons exist. An early season, which is covered in this example, and a summer monsoon season. In this year the summer monsoon began July 1 and, for simplicity is not included in the example).

Data quality is initially examined by plotting observations on scattergrams using SPSS subprogram SCATTERGRAM in the following manner:

```

RUN NAME          INITIAL SCATTERGRAMS - EXAMPLE PROGRAM 1
VARIABLE LIST     YR,MO,KY,SW,TX,TN,ADP,MDP,PA,YTX,YTN,YADP,YMDP
INPUT MEDIUM     FIXED TAPE
N OF CASES        80
INPUT FORMAT      (2X,3F2.0,F1.0,2F3.0,3F4.0,2F3.0,2F4.0)
MISSING VALUES   SW TO YMDP (-99)
SCATTERGRAM       YTX WITH TX,SW, PA/YTN WITH TN/YADP WITH ADP/YMDP WITH MDP
OPTIONS           7
READ INPUT DATA
FINISH

```

This produces six plots with all the observations on at least one of them.

On examining these graphs a minimum temperature of 5⁰F at Flagstaff is noted in June. (The scattergrams will be used again in the second phase). By comparing that value with other surrounding dates and maximum temperature for that day, it appears to be an error. Also there are two apparent errors of

YADP as well as some questionable points. It is decided that if YADP and ADP differ by more than 25°F then they are defined to be erroneous. This eliminates three values of YADP. The following additions are made to the previous program:

```

      .
      .
IF      (TN LT 6) TN = -99
COMPUTE TDIFF = ABS (YADP-ADP)
IF      (TDIFF GT 25) YADP = 99
SCATTERGRAM as given
      .
      .

```

These cards do not change the value on the data file which we wish to preserve in case the correction needs to be changed later. The data set now looks acceptable and calculation can begin.

3. Term Development

This phase develops and finalizes the form of the independent variables which are considered for inclusion in the regression equation. There are three major steps in this phase. In the first phase, consideration is given to transformations of individual independent variables (base station observations). The second and third steps develop new independent variables from combinations of existing independent variables. These steps are performed in order to account for nonadditive relationships between variables and to add to the forms of the variables which can be used in the regression equations. (The second and third steps result in some potential difficulties which are treated in the third developmental phase). As will be seen, the three steps are not entirely independent of each other.

3.1 Observations and Transformations

Observations that can be used are from the permanent fire weather station and are generally limited to those stored in the National Fire Weather Data Library (Furman and Brink, 1975). Only those that are pertinent, reliable

and frequent should be used. Most of the observations can be used directly but some may require transformation. The State of Weather Code is one of these. Values 0 through 3 denote classes of cloud cover if there is no precipitation, and 4 through 9 code the nature of precipitation. There are a number of transformations possible but the one found most useful treats only the cloud cover codes and inverts their order so that they are inversely proportional to cloud cover. If SW is greater than 3, it is set to 3 and that value is then subtracted from 4. In SPSS statements:

```
IF          (SW GT 3) SW = 3
COMPUTE     SW = 4 - SW
```

For some variables a log or square root type of transformation may be desired. These are variables which have a curved relationship illustrated by the scattergrams of the data which were plotted previously. Without a physical reason, the decision to do this type of transformation (and which one) is a subjective one based on inspection of scattergrams. In dry climates the dew point can go below zero so a constant must be added to it before some transformations are made. For instance the square root of the afternoon dew point may be coded.

```
COMPUTE     ADPSQRT = SQRT(ADP+20)
```

where -19 is the lowest dew point.

3.2 Combined Terms

Testing has shown that certain combinations of independent variables can be beneficial to the final results. In the next paragraph three of these are introduced which are differences between independent variables. These new variables cannot actually add new information to the regression equations: they only restructure the form of the variables. However, this gives

the stepwise variable selection program a choice of forms. Experience has been that these terms are frequently chosen. (On the negative side, a problem of multicollinearity is introduced. This is discussed in Section 4).

Having transformed relative humidity to dew point it might be useful to have a measure of relative saturation in the set of independent variables. The observed relative humidity values at the base station can be used but it is more convenient to use dew point depressions (the difference between the dew point and the temperature). In SPSS code the afternoon and morning dew point depressions are computed with

```
COMPUTE      ADPD = TX-ADP
COMPUTE      MDPD = TN-MDP
```

The temperature range is another variable that conceivably may be useful. Cloud cover that is coded by SW is only valid at observation time while temperature range is related to night time and early day cloud cover (and other effects). This variable is generated by the statement

```
COMPUTE      TR = TX-TN
```

These three new variables are added to the example. Certainly the number of variables that can be created in this way are not limited to these three but these have been found to be most useful.

3.3 Interactive Terms from Correlation Considerations

Implicit in the form of a regression equation is the assumption that the effects of the independent variables are additive. This is not always the case. One way to handle this problem is through log or similar transformations as discussed previously. Another way to treat non-additive effects are through interactive terms. Usually, without reason to do otherwise,

interactive terms are of multiplicative form. Two or more independent variables are multiplied by each other. Obviously with several (nine in the example) independent variables already available, the number of interactive variables possible is huge. Therefore two limitations which have been successful in testing are imposed. First only two independent variables are multiplied to form interactive terms. Secondly these two variables must be correlated with each other.

A useful tool for developing interactive terms is the SPSS subprogram PEARSON CORR. Correlation coefficients are computed between independent variables and between the noninteractive terms and the dependent variables.

For the example this is done with the following SPSS program:

```

RUN NAME          PEARSON CORR - EXAMPLE PROGRAM 2
VARIABLE LIST     YR,MO,DY,SW,TX,ADP,MDP,PA,YTX,VTN,YADP,YMDP
INPUT MEDIUM     TAPE
INPUT FORMAT      FIXED (1X,3F2.0,F1.0,2F3.0,3F4.0,2F3.0,2F4.0)
N OF CASES        80
MISSING VALUES   SW TO MDYP (-99)
IF                (TN LT 6) TN = -99
COMPUTE           ADPD = TX-ADP
COMPUTE           MDPD = TN-MDP
COMPUTE           TR = TX-TN
COMPUTE           PA = .01*PA (optional)
IF                (SW GT 3) SW = 3
COMPUTE           SW = 4-SW
ASSIGN MISSING    ADPD TO TR (-99)
PEARSON CORR      YTX,SW,TX,TN,ADP,MDP,PA,ADPD,MDPD,TR/
                  YTN,SW,TX,TN,ADP,MDP,PA,ADPD,MDPD,TR/
                  YADP,SW,TX,TN,ADP,MDP,PA,ADPD,MDPD,TR/
                  YMDP,SW,TX,TN,ADP,MDP,PA,ADPD,MDPD,TR
OPTIONS           2,6 (optional)
STATISTICS        1 (optional)
READ INPUT DATA
FINISH

```

Table 1 contains a synthesis of the correlation matrices that result from this program. (The Table is not in SPSS output form for reasons of clarity and space saving).

Table 1

Correlation Coefficients From Example Program 2

		SW	TX	TN	ADP	MDP	PA	ADPD	MDPD	TR
1	YTX	.30	.91	.56	.26	.18	-.44	.62	.39	.67
2	YTN	-.44	.63	.82	.59	.60	-.07	.08	.17	-.03
3	YADP	-.47	.06	.59	.69	.60	.24	-.52	-.20	-.46
4	YMDP	-.63	.17	.70	.66	.72	.15	-.44	-.24	-.48
5	SW	1	.16	-.26	-.57	-.52	-.48	.61	.37	.50
6	TX		1	.71	.35	.29	-.36	.62	.42	.62
7	TN			1	.62	.61	-.07	.14	.35	-.11
8	ADP				1	.86	.22	-.51	-.35	-.19
9	MDP					1	.23	-.44	-.53	-.27
10	PA						1	-.52	-.34	-.44
11	ADPD							1	.68	.73
12	MDPD								1	.20
13	TR									1

The fifth through thirteenth rows of Table 1 contain the correlation coefficients (r) between the nine independent variables that have been developed to this point. Using these coefficients for guidance we wish to select pairs of variables to form multiplicative terms. There are nine pairs of variables (rows 5-13) with $r > .6$. This seems to be a large but not inconvenient number of new terms to allow for possible inclusion in the regression equations. Only pairs of independent variables with $r > .6$ are multiplied to form interactive terms. From the correlations of Table 1 the following interactive terms are now defined; SWADPD(SW*ADPD), TXTN(TX*TN), TXADPD, TXTR, TNADP, TNMDP, ADPMDP, ADPDMDPD, and ADPDTR.

Recall that SW is an ordinal level variable. It has only one of four possible values and has a range of only four units. These characteristics tend to limit its variance properties and its use as a predictor. Furthermore, interactive terms allow the effects of a variable to change with its level. It may be useful to allow the effects of some variables to change with cloud cover codes. The arbitrary level of $r > .6$ is then relaxed to $r > .5$ for SW so, from row 5 of Table 1, the terms SWADP and SWMDP seem appropriate. Remember that in this phase of the regression procedure variables are being developed for possible inclusion in regression equations. A variable that is not defined in this phase cannot be considered later but the decision to not use a variable can and will be made later. It is preferable to err, within limits, on the side of too many variables. (As for the combined terms, a potential for multicollinearity from these terms exists. This is treated in Section 4).

3.4 Nature of the Independent Variable Set

There are now twenty possible independent variables in the example. From this set, regression equations to estimate dependent variables are to be generated in the next phase. Before doing so, a look at what is and is not in the variable set is worthwhile.

First of all, the variable list contains only measurements that are related to certain manifestations of physical processes. The complex physics of climatic variations are not explicitly involved. There is no treatment of or any feasible observations of such factors as atmospheric stability, air mass type and surface albedo. Hopefully these factors are incorporated adequately in the observations which can be used. Proper partitioning of seasons is also important in separating some of the long term and/or slowly changing influences. The interactive variables are difficult to interpret physically. Indeed, they may have little physical meaning. However, they do tend to add to the predictive capability of the regression equations that are to be determined and should not be excluded. Also, note that physiographic influences such as elevation and aspect do not appear in the independent variable set. Between any two sites the differences of these factors are constants and cannot appear explicitly in regression equations. Differences in physiography are included implicitly in the observation differences of the two weather stations.

The discussion so far has considered the development of the independent variable set as being a series of decisions. Admittedly, a novice at this procedure may not feel ready to make all of these decisions. On the other hand, someone skilled in meteorology and regression may have additional ideas. The example that is presented is a result of repetitive testing and much

trial and error work on four years of data from nine remote stations in two areas (see Part 1). It is expected that following the example as close as possible would yield acceptable results. That is, the example represents the best of what works in testing the procedure and as such presents what are found to be the best of many options. But the example can be used as a starting point for those who wish to try additional steps for their situations. Once the data file is ready, it is surprisingly easy to modify the SPSS programs and to experiment with new variables. In fact, users who have the time and skills are encouraged to do so.

Testing has shown that the procedure for deriving variables as given in the example extracts about all the useful information from the observations. For the test data, additional sophistication or more detailed treatments have not resulted in improvements of the final product. In fact there are probably too many variables in the variable set; several of them are redundant in their explanation of changes of the dependent variable. In the next phase, we give the task of selecting the best of these to SPSS software.

4. Determining Regression Equations

The goals of this phase are to let the computer select regression equations (one for each dependent variable) and then to examine their fit to the observations. An explanation of the stepwise multivariate regression procedure is given in Nie et. al. (1975) and is treated here much like a "black box." However, the results must be examined and a determination must be made on whether or not the output can be used as the final equations. Each of these steps are discussed separately. (In the following text, unstandardized regression coefficients are used. The coefficients may have some

bias which can be reduced by using standardized coefficients. This refinement of the regression procedure is considered as optional and is discussed in Appendix D).

4.1 Calculations

SPSS subprogram REGRESSION is used in this step. With a couple of exceptions all control parameters are permitted to have default values (see Nie, et. al. (1975) and updates as appropriate). The regression procedure statements have the form

```
REGRESSION      METHOD      = regression method/VARIABLES = variable list/
                REGRESSION = design statement/RESIDUALS/
                REGRESSION = design statement/RESIDUALS/
                :
                :
```

The regression method here is STEPWISE, the variable list for the example is the set of four dependent and twenty independent variables, the design statement is to be discussed and the word RESIDUALS (or RESID) enables a residual analysis. The design statement has the form

```
REGRESSION = dependent variable (NSTEPS,FIN,TOL,FOUT)
            WITH indep. vars.
```

The dependent variable is the one for which estimates are desired, the parameters in parentheses are control parameters and the independent variables are those of the previous phase.

For the example, the regression subprogram portion of the SPSS program is:

```
REGRESSION      METHOD = STEPWISE/VARIABLES = SW TO YMDP,ADPD TO SWMDP
                REGRESSION = YTX(*,1.0,*,1.0) WITH SW TO PA,ADPD TO SWMDP
                /RESID/
                REGRESSION = YTN(*,1.0,*,1.0) WITH SW TO PA,ADPD TO SWMDP
```



```

/RESID/
REGRESSION = YADP(*,1.0,*,1.0) WITH SW TO PA,ADPD TO SWMDP
/RESID/
REGRESSION = YMDP(*,1.0,*,1.0) WITH SW TO PA,ADPD TO SWMDP
/RESID/
OPTIONS      6          (optional)
STATISTICS   4.5        (needed, others optional)

```

There is one regression design statement for each dependent variable. Statistics 4 and 5 are for a residual analysis in the next step. The above statements permit the number of steps (NSTEPS) and tolerance levels (TOL) to be free. The statistical F ratios, FIN and FOUT, are criteria for including and excluding terms in the regression equation. Default limits are for more generous than are needed. Values of 1 to 2 have been found to be the most useful but these are arbitrary choices. In tests, setting FIN = FOUT = 1 resulted in equations with typically four or five terms while default levels permitted 10-15 terms - most of which were useless. The full regression program for the example is given in Appendix C.

There are four blocks output for each variable. The first describes the regression equation after the final step. It contains a list of each of the terms selected, the coefficient for each term and a table of descriptive statistics for the equations fit to the observed values. The second block consists of a summary table which shows which variable is entered or removed from the equation at each step and the resulting changes of some descriptive statistics. An extraction for the YTX variable of the example of these outputs is given in Table 2. The third and fourth portions result from the STATISTICS selections and are discussed in the next section.

The regression equation is of the general form

$$Y' = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

where $B_0, B_1, B_2 \dots B_n$ are regression coefficients and $X_1, X_2 \dots X_n$ are independent

Table 2

Some Output From The REGRESSION Subprogram For The Equation For YTX

DEPENDENT VARIABLE.. YTX

FINAL STEP.

MULTIPLE R	.910	ANAL. OF VAR.	DF	SUM OF SQ	MEAN SQ	F	SIG
R SQUARE	.873	REGRESSION	5	2940.77	588.14	83.8	.000
ADJUSTED R SQUARE	.866	RESIDUAL	74	622.71	8.42		
STANDARD ERROR OF ESTIMATE	2.960	COEFF. OF VAR.	3.4 PCT				

Variables In Equation

Variables Not In Equation

VARIABLE	B	STD. ERROR OF B	F/SIG	(the fifteen remaining values are listed)
ADPD	.5608 E-2	.46 E-2	14.67/.000	
TXTN	.1086 E-1	.75 E-3	207.9/0	
SWADP	.993	.622	2.556/.115	
TNADP	.458 E-2	.371 E-2	1.528/.221	
PA	-24.087	6.281	1.790/.186	
(Constant)	39.980	2.755	210.5/0	

STEP	VARIABLE		F TO ENTER		R SQUARE	
	ENTERED	REMOVED	OR REMOVE	SIGNIFICANCE	R SQUARE	CHANGE
1	TX		184.04	.0	.742	.742
2	ADPD		12.77	.001	.785	.043
3	TXTN		13.22	.001	.823	.038
4	SWADP		14.03	.000	.856	.033
5		TX	.36	.552	.855	-.001
6	TNADP		2.00	.162	.860	.005
7	PA		1.51	.223	.866	.006

variables for n terms. For the example equation as given by Table 2

$$YTXE = 39.980 + .0056*ADPD + .0109*TXTN + .933*SWADP + .004588*TNADP -24.087*PA$$

The variable TX is entered on the first step but is removed in the fifth step.

About 87% of the variance of YTX is explained and about two thirds of the residuals (YTX-YTXE) are within about $\pm 3^\circ F$. All values in the output are explained in Nie et. al. (1975). The three other equations are taken from the output in a similar manner.

A problem of multicollinearity can exist when two or more independent variables are highly correlated. The introduction of combined terms and interactive terms increases the probability of experiencing this kind of problem. Multicollinearity problems result in unreliable regression coefficients so it is important to check that the output coefficients are correct. (On Control Data Corp. computers a "perfect" equation is given; $R^2 = 1$. This is false and due to multicollinearity.) The coefficients should be checked by hand using the mean value of each of the selected independent terms. The sum of the constant and the means of independent terms multiplied by the appropriate coefficients should give the mean of the dependent variable (i.e. $\bar{Y} = B_0 + \sum_{i=1}^n B_i \bar{X}_i$). If this is not true (and the hand calculations are correct) then corrective actions must be taken and the regression program must be rerun. These are several ways to correct this problem. In addition to those given in SPSS documentation (Nie et. al., 1975; p 341) two other choices are possible. One is to use double precision variables if they are available on the SPSS version being used. The second is to refer to the summary table and stop the REGRESSION subprogram at an earlier step (by specifying a value for NSTEP) that would still give good results. With any

choice the new coefficients should be checked again for multicollinearity problems.

The statistics presented in the final step and summary table are used to evaluate the performance of the regression equation for the year of observation. If the equation is not good at this point then it is not useful to apply it to historical records. The meaning of each of the statistics are given in Nie et. al. (1975). There are many ways to judge if an equation is satisfactory. Here we set a minimum requirement that the coefficient of variation must be less than 30%.

4.2 Examination of the Residuals

Residuals are defined as the difference of the observed and estimated dependent variables (YTX-TYXE for example). It is desired that all residuals be small and their signs be randomly ordered in time sequence. Using daily data it is too much to require random order of signs. Meteorological events take place on the order of two or three days so small periods of residuals of the same sign must be expected. What should not be accepted are large trends in the residuals over the season or long periods of "runs" of residual signs.

A trend indicates that the statistical relationship between the two stations is changing with time. This may be due to improper partitioning of seasons. In the early development of the regression procedure residual trends (sometimes sign shifts) were observed for a location in Idaho. Typically, residuals went from generally negative to positive around the first third of July. The cause could not be determined with certainty but the sign change dates coincided roughly with the time of the curing of grasses. Residual trends were eliminated or greatly reduced by forming two seasons - before and after July 7.

If a run of positive or negative residuals is found at either end of the season, it might be concluded that they do not belong to the season. If so these days should be removed from the data set and the programs should be rerun. Similarly it would be valid to exclude data taken during an unusually late snowstorm because they wouldn't represent the fire season. If major changes of the data set or the season are made, then both the second and third phases must be redone. With minor changes only rerunning the REGRESSION subprogram is needed.

Examination of residuals is enabled by using STATISTICS 4 and 5 with the REGRESSION subprogram. These result in the printing of a day by day plot of residuals, a list of daily dependent variable observations and estimates and a statistical summary of the residuals. A sample statistical summary is given in Table 3. There does not seem to be a unique set of objective criteria for deciding if the results are satisfactory in this application. Minimally, the numbers of positive or negative residuals should not be less than a third of the number of days used. The number of runs of signs should not be less than half of the expected number of runs. The reader is referred to SPSS documentation for explanation of the statistics in the residual summary table. The magnitudes of the individual residuals can be examined in the residual plot. The plot may be more useful in looking for trends, checking for appropriate season beginning and ending dates and finding unusual events that the user may want to eliminate from the analysis (i.e. a late snowstorm).

If the user determines that the regression equations are satisfactory, then they can be applied to historical records.

Table 3

Example Statistics For Residuals

VON NEUMAN RATIO	1.666	DURBAN-WATSON TEST	1.645
NUMBER OF POSITIVE RESIDUALS	37.		
NUMBER OF NEGATIVE RESIDUALS	43.		
NUMBER OF RUNS OF SIGNS	40.		
EXPECTED NUMBER OF RUNS OF SIGNS	41.		
EXPECTED S.D. OF RUN DISTRIBUTION	4.418		
UNIT NORMAL DEVIATE -			
$Z = (\text{Expected-Observed})/\text{S.D.}$	-0.06224		
PROBABILITY OF OBTAINING .GE. $ABS(Z)$.475		

5. Application

Upon completion of the regression equation development phases there is one equation for each variable to be estimated and for each season. Estimation of fire climate at the remote site is now possible by applying the regression equations to the long term records of the base station.

The first step is to copy the records desired from the National Fire Weather Data Library to the users own file. These records should be spot checked and/or plotted in time series to determine if there are periods of incomplete data. For efficiency, any records (days) that do not have complete observations of the independent variables should be deleted from the file. Beginning and ending dates of each season for each year should be determined. Optimally, equations developed from a season should be applied to all days of the season without regard to calendar date. If, as in the example study area, there is one set of equations for the early summer and another set for the summer monsoon, then the second set should

not be applied until the monsoon begins for each year. It is often more convenient to choose an average calendar date for season change but the user should be aware if doing so would result in unacceptable error.

Estimates of dependent variables can now be made (with a user supplied computer program). In the process of estimating humidity variables dew points must be calculated before the estimate is made. Estimates of relative humidity can be made from the predicted values of temperatures and dew points. Equations for performing these transformations are given in Appendix B. The regression equations are then applied to the historical records of the base station to estimate the climate in terms of the regressed variables at the remote site. For the year that the remote station is operated no estimates need be made; the observations are used. (However by making those estimates with the user's program and comparing them to those from the residual analysis, a convenient check of the user's program is obtained.)

Fire models require some input meteorological variables in addition to those which can be estimated with the regression equation (SW, PA etc.). At present, the best approximation for these variables are the observations at the base station. These data should be copied from the National Fire Weather Data Library files onto the same file as the estimated variables. The combined file is then used as input to the fire models.

6. Outline Review of Procedures

I. Equation Development

A. Phase 1 - Set-up

1. Remote station site selection
2. Season(s) definition
3. Base station selection
4. Measurements
5. Data reduction
6. Create file
7. Quality control screen
 - a. plots (SCATTERGRAM)
 - b. corrections
8. Finalize file

B. Phase 2 - Term Development

1. Observations
 - a. direct
 - b. transformed (ln, SQRT, etc., SCATTERGRAM)
2. Combined terms
3. Interactive terms
 - a. Correlation matrix (PEARSON CORR)
 - b. Generate new terms from matrix
 - c. other considerations (i.e. ordinal level variables)

C. Phase 3 - Equation Determination and Evaluation

1. Regression design
 - a. Control parameter specifications
 - b. Regression design statements

- c. Select OPTIONS
- d. Select STATISTICS
- 2. Examine coefficients for multicollinearity problems
- 3. Final step statistics check
 - a. $R^2 > .5$ (except wind speed)
 - b. Coefficient of variation $< 30\%$.
- 4. Residual Examination
 - a. plots
 - i. trends
 - ii. events
 - iii. magnitudes
 - b. statistical summary
 - i. numbers of positive or negative residuals
 - ii. number of runs of signs
- 5. Decision of Adequacy
 - a. Go back
 - b. Go forward

II. Application

- A. Create file of base station data
 - 1. Check for missing data
 - 2. Check for unreliable or unusable data
 - 3. Modify file as needed
- B. Season changes
 - 1. Individual years
 - 2. Average dates
 - 3. Adjust as necessary
- C. Make estimates
- D. Finalize file of remote station climate information

References

- Barrows, J. S. and L. K. Balick. (1977). Meteorologic Impacts of Forest Management Activities on Wildfire Potential. Interim Progress Report on file at the Rocky Mountain Forest and Range Experiment Station. USDA Forest Service, Ft. Collins, CO.
- Draper, N. R. and H. Smith. (1966). Applied Regression Analysis. John Wiley and Sons, Inc., New York, N. Y. 407 pp.
- Furman, R. W. and G. E. Brink. (1975). The National Fire Weather Data Library: What it is and how to use it. USDA Forest Service, General Technical Report RM-19, Rocky Mountain Forest and Range Experiment Station, Ft. Collins, CO. 8 p.
- Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner and D. H. Bent. (1975). Statistical Package for the Social Sciences. McGraw-Hill Book Company, New York, N. Y. 675 pp.

Appendix A

Variable Nomenclature

In this report a set of symbols have been used which assign short, computer compatible names to variables. The following set of symbols form the basis for the nomenclature used:

TX	daily maximum temperature (°F)
TX	daily minimum temperature (°F)
ADP	afternoon dew point (°F)
MDP	precipitation amount (°F)
SW	state of weather code
PA	precipitation amount (inches)

The first four of the above can exist as independent variable observations, dependent variable observations or dependent variable estimates. To distinguish these forms, a Y is placed in front of the variable name for dependent variables and an E follows the name if it is an estimate. Thus the maximum temperature observed at the remote (dependent) site is YTX and the estimate is YTXE.

Fourteen additional independent variables are generated in terms of the above six variables. These are defined as

ADPD = TX-ADP	(afternoon dew point depression)
MDPD = TX-MDP	(morning dew point depression)
TR = TX-TN	(temperature range)
SWADPD = SW * ADPD	(this and the following ten variables are multiplicative interactive terms)
TXTN = TX * TN	
TXADPD = TX * ADPD	
TXTR = TX * TR	
TNADP = TN * ADP	
TNMDP = TN * MDP	
ADPMDP = ADP * MDP	
ADPDMDPD = ADPD * MDPD	
ADPTR = ADP * TR	
SWADP = SW * ADP	
SWMDP = SW * MDP	

Appendix B uses the following variable names in addition to those defined above

AES	afternoon saturation (water) vapor pressure in mb.
MES	morning saturation vapor pressure in mb.
AE	afternoon vapor pressure in mb.
ME	morning vapor pressure in mb.
RHN	minimum (afternoon) relative humidity in percent
RHX	maximum (morning) relative humidity in percent
T	an intermediate temperature value in °K
DP	an intermediate dew point in °K

Appendix B

Relative Humidity - Dew Point Conversions

The following equations are used to calculate dew points from temperature and relative humidity: for afternoon dew point:

$$T = .55556 * (TX-32) + 273.16 \text{ (converts TX from } ^\circ\text{F to } ^\circ\text{K)}$$

$$AES = 6.108 \exp [(A*(T-273.16))/(T-B)] \quad (A = 17.269, B = 35.86)$$

$$AE = .01 * RHN * AES$$

$$ADP = (5352.2/(21.4-\ln(AE)))-273.16 \text{ (in } ^\circ\text{F)}$$

for morning dew point:

$$T = .55556 * (TN-32) + 273.16$$

$$MES = 6.108 \exp [(A*(T-273.16))/(T-B)]$$

$$ME = .01 * RHX * MES$$

$$MDP = (5352.2/(21.4-\ln(ME))) - 273.16$$

More details are given in Appendix A, Part 1 of this report.

To convert dew point to relative humidity, the following equations are used:

for RHN;

$$T = .55556 * (TX-32) + 273.16$$

$$DP = .55556 * (ADP-32) + 273.16$$

$$AES = 6.108 \exp ((A*(T-273.16))/(T-B))$$

$$AE = 6.108 \exp ((A*(T-273.16))/(T-B))$$

$$RHN = (AE/AES) * 100 \text{ in percent}$$

(A and B are constants defined above)

for RHX;

$$T = .55556 * (TN-32) + 273.16$$

$$DP = .55556 * (MDP-32) + 273.16$$

$$MES = 6.108 \exp ((A*(T-273.16))/(T-B))$$

$$ME = 6.108 \exp ((A*(DP-273.16))/(T-B))$$

$$RHX = (ME/MES) * 100$$

Note that the coefficients 6.108 divide out and need not be included in the calculations. For the derivation of these relationships see Appendix A, Part 1.

Appendix C

Example REGRESSION Program

Because of its length the entire SPSS program to compute the regression equations and associated statistic is presented here.

```

RUN NAME          REGRESSION EXAMPLE PROGRAM 3
VARIABLE LIST     SW,TX,TN,ADP,MDP,PA,YTX,YTN,YADP,YMDP
INPUT MEDIUM      TAPE
INPUT FORMAT      FIXED(2X,3F2.0,F1.0,2F3.0,3F4.0,2F3.0,2F4.0)
N OF CASES        80
MISSING VALUES   SW TO YMDP (-99)
COMPUTE           TDIFF = ABS (YADP-ADP)
IF                (TDIFF GT 25) YADP = -99
IF                (TN LT 6) TN = -99
COMPUTE           PA = 0.01*PA
IF                (SW GT 3) SW = 3
COMPUTE           SW = 4-SW
COMPUTE           ADPD = TX-ADP
COMPUTE           MDPD = TN-MDP
COMPUTE           TR = TX-TN
COMPUTE           SWADPD = SW *ADPD
COMPUTE           TXTN = TX * TN
COMPUTE           TXADPD = TX * ADPD
COMPUTE           TXTR = TX * TR
COMPUTE           TNADP = TN * ADP
COMPUTE           TNMDP = TN * MDP
COMPUTE           ADPMDP = ADP * MDP
COMPUTE           ADPDMDPD = ADPD * MDPD
COMPUTE           ADPDTR = ADPD * TR
COMPUTE           SWADP = SW * ADP
COMPUTE           SWMDP = SW * MDP
ASSIGN MISSING    ADPD TO SWMDP (-99)
ECOLOGY           ON
REGRESSION        METHOD = STEPWISE/VARIABLES = SW TO YMDP,ADPD TO SWMDP
                  REGRESSION = VTX(*,1.0,*,1.0) WITH SW TO PA ADPD TO SWMDP
                  /RESID/
                  REGRESSION = YTN(*,1.0*,1.0) WITH SW TO PA ADPD TO SWMDP
                  /RESID/
                  REGRESSION = YADP(*,1.0,*,1.0) WITH SW TO PA ADPD TO SWMDP
                  /RESID/
                  REGRESSION = YMDP(*,1.0,*,1.0) WITH SW TO PA ADPD TO SWMDP
                  /RESID/
OPTIONS           6
STATISTICS        4, 5
READ INPUT DATA
FINISH

```

Appendix D

Use of Standardized Regression Coefficients

Regression coefficients which are computed are estimates of some hypothetical "true" value. This is analogous to the difference between a sample mean and a population mean. In regression analysis the regression coefficients are unbiased if the statistically "correct" regression model is chosen (Draper and Smith, 1966). If the estimates are biased, then the degree of bias is, in part, dependent on the magnitudes of the independent variables. There is an advantage then to standardizing all variables so that each has a mean of 0 and a standard deviations are 1 unit. This is done by subtracting the mean from each measurement and dividing that difference by the standard deviation $((X_{ij} - \bar{X}_j)/s_j$ for all days (i) and all variables (j)). The regression equation has the same form as before except $B_0 = 0$.

SPSS routinely outputs standardized regression coefficients although these are not shown in the example presented in the text. When applying the equation the variables must already be transformed and a reverse transformation must be applied to the estimates.

In both Part 1 and Part 2 the regression procedure is developed and tested without standardization. This was done in order to keep the procedure as simple as possible. (Also, in Part 1, testing was done on two hundred and sixteen equations and keeping track of a mean and standard deviation for each proved to be very unpleasant.) Rigorously, for the transformation to yield variables with a mean of 0 and a standard deviation

of 1, the measurements must have a normal distribution. This is not true for several of the variables that are used, maximum temperature, precipitation variables, and state of the weather code. (The transformation of these variables still tends to reduce bias.) In this context, the authors consider the use of standardized data and coefficients as an optional refinement of the regression procedure.